Identification, alignement, et traductions des adjectifs relationnels en corpus comparables

Rima Harastani, Beatrice Daille, Emmanuel Morin

Identification, alignement, et traductions des adjectifs relationnels en corpus comparables

Rima Harastani[1] Beatrice Daille[1] Emmanuel Morin[1]
(1) LINA UMR CNRS 6241, 2 Chemin de la Houssinière 44300 Nantes
{rima.harastani,beatrice.daille,emmanuel.morin}@univ-nantes.fr

RÉSUMÉ

Dans cet article, nous extrayons des adjectifs relationnels français et nous les alignons automatiquement avec les noms dont ils sont dérivés en utilisant un corpus monolingue. Les alignements adjectif-nom seront ensuite utilisés dans la traduction compositionelle des termes complexes de la forme [N AdjR] à partir d'un corpus comparable français-anglais. Un nouveau terme [N N′] (ex. cancer du poumon) sera obtenu en remplaçant l'adjectif relationnel *AdjR* (ex. pulmonaire) dans [N AdjR] (ex. cancer pulmonaire) par le nom N' (ex. poumon) avec lequel il est aligné. Si aucune traduction n'est proposée pour [N AdjR], nous considérons que ses traduction(s) sont équivalentes à celle(s) de sa paraphrase [N N′]. Nous expérimentons avec un corpus comparable dans le domaine de cancer du sein, et nous obtenons des alignements adjectif-nom qui aident à traduire des termes complexes de la forme [N AdjR] vers l'anglais avec une précision de 86 %.

ABSTRACT

Identification, Alignment, and Tranlsation of Relational Adjectives from Comparable Corpora

In this paper, we extract French relational adjectives and automatically align them with the nouns they are derived from by using a monolingual corpus. The obtained adjective-noun alignments are then used in the compositional translation of compound nouns of the form [N ADJR] with a French-English comparable corpora. A new term [N N′] (eg, cancer du poumon) is obtained by replacing the relational adjective *AdjR* (eg, pulmonaire) in [N AdjR] (eg, cancer pulmonaire) by its corresponding N' (eg, poumon). If no translation(s) are obtained for [N AdjR], we consider the one(s) obtained for its paraphrase [N N′]. We experiment with a comparable corpora in the field of breast cancer, and we get adjective-noun alignments that help in translating French compound nouns of the form [N AdjR] to English with a precision of 86%.

MOTS-CLÉS : Adjectifs relationnels, Corpus comparables, Méthode compositionnelle, Termes complexes.

KEYWORDS: Relational adjectives, Comparable corpora, Compositional method, Complex terms.

1 Introduction

Les termes complexes sont des termes qui se composent de plus d'un mot. La plupart de ces termes possèdent une propriété compositionnelle, c'est-à-dire que la signification de l'ensemble peut être appréhendée par la signification des parties. Ainsi, certaines approches ont été proposées pour traduire des termes complexes en fonction de cette propriété (voir Baldwin et Tanaka (2004)). Elles consistent à traduire un terme complexe mot à mot à l'aide d'un dictionnaire bilingue. Ensuite, elles combinent ces traductions individuelles selon des formes appropriées pour produire des traductions candidates du terme complexe. Les traductions candidates sont ensuite cherchées dans un corpus comparable [1] avant d'être considérées comme correctes. Les corpus comparables ont été utilisés avec succès dans la tâche de l'alignement de termes par de nombreuses approches (Rapp, 1995; Baldwin et Tanaka, 2004) en raison de leur plus grande disponibilité par rapport aux corpus parallèles [2] (Bowker et Pearson, 2002). Ainsi, pour traduire compositionnellement le terme français "gestion clinique" en anglais, on peut traduire "gestion"

1. des textes multilingues qui appartiennent au même domaine
2. textes multilingues qui sont des traductions mutuelles

par "management" et "clinique" par "clinical", puis rassembler ces traduction sous la forme [A N] (*A* et *N* signifient respectivement adjectif et nom) afin d'obtenir une traduction candidate "clinical management".

Nous nous intéressons aux termes complexes de la forme [N AdjR] (*Ad jR* désigne un adjectif relationnel), ex. cancer pulmonaire. En effet, ces termes peuvent être traduits compositionnellement dans une autre langue par des termes de la forme [N N] (ex. "cancer pulmonaire" est traduit en anglais par "lung cancer"), voir figure 1. Si le substantif "lung" n'est pas une traduction de l'adjectif "pulmonaire" dans le dictionnaire, le lien entre "pulmonaire" et "lung" peut être établi via le substantif "poumon" dont "pulmonaire" est le dérivé et "lung" est la traduction. Ainsi, nous pouvons traduire "cancer pulmonaire" par "lung cancer" en passant par la paraphrase "cancer du poumon". Cette piste a été déjà explorée par Morin et Daille (2010) à l'aide des règles définies qui relient un adjectif relationnel avec son nom. Nous avons pour objectif dans ce travail (a) d'extraire des adjectifs relationnels automatiquement du corpus ; (b) d'établir un lien entre un adjectif relationnel extrait et le nom dont il est dérivé automatiquement ; (c) d'étudier l'influence des propriétés des adjectifs extraits et les alignements adjectif-nom sur la traduction compositionnelle des termes [N AdjR].

Après une présentation de la classe d'adjectifs relationnels et les problèmes liés à son identification en section 2, nous développons dans la section 3 une approche qui nous permet d'extraire automatiquement des adjectifs relationnels d'un corpus français. Ensuite, nous proposons une approche en section 4 afin de relier un adjectif relationnel (extrait précédemment du corpus) à un nom existant dans un dictionnaire bilingue et dans le corpus. Si la plupart des adjectifs relationnels sont dérivés par suffixation à partir de noms populaires (ex. cancéreux / cancer), il y en a d'autres qui sont construits à partir de racines supplétives des noms (ex. médullaire / moelle). Nous traitons ces deux cas séparément : (a) **adjectif relationnel commun** : nous supposons qu'un adjectif relationnel partage un certain nombre de lettres avec son nom de base, et que l'ordre des lettres est conservé. Ainsi, un score entre un adjectif relationnel et chaque nom dans un dictionnaire (et qui existe dans le corpus) sera obtenu en fonction de la similarité de lettres par l'approche décrite en section 4.1, nous exploitons ensuite le contexte afin que ce score soit plus représentatif en section 4.1.1 ; (b) **adjectif relationnel savant** : nous vérifions si un adjectif relationnel peut être relié avec un nom à l'aide d'une racine supplétive en appliquant l'approche expliquée en section 4.1.3. En section 5, nous utilisons les alignements obtenus par l'approche d'alignement adjectif-nom dans la traduction compositionnelle par paraphrase des termes [N AdjR]. Enfin, nous évaluons en section 6 les approches que nous proposons et nous concluons dans la section 7.

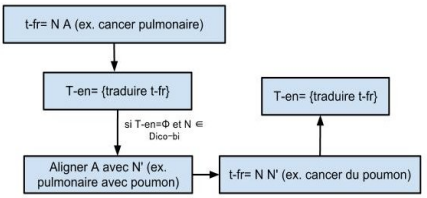

FIGURE 1 – Traduction par paraphrase (où t-fr est le terme français, T-en est l'ensemble des traductions anglaises et Dico-bi est le dictionnaire bilingue français-anglais).

2 Adjectifs relationnels

Dans cette section, nous présentons la classe des adjectifs relationnels et ses propriétés, ainsi que des travaux qui se sont intéressés à l'identification de cette classe et les problèmes liés à cette identification.

2.1 Définition et propriétés

D'après Dubois et Dubois-Charlier (1999, p. 128), "*un adjectif relationnel est issu d'une relative, où <de N> est caractérisé par l'absence de déterminant ; cette relative se branche directement sur l'antécédent*

auquel elle se rapporte, et l'ensemble formé du nom et de l'adjectif nominal suffixé forme un nom composé". Exemple : ce corps chimique est l'acide qui est <de nitre> ; ce corps chimique est l'acide nitrique.

Les adjectifs relationnels sont des adjectifs dénominaux (adjectifs construits sur des bases nominales), à ne pas confondre avec les adjectifs déverbaux qui sont dérivés d'un verbe par des suffixes tels que *-able, -ible, -ile, -ant, etc*, (ex. dégradable/ de dégrader). Alors qu'un adjectif dit "qualificatif" (*AdjQ*) peut aussi être construit sur une base nominale, la relation [N AdjQ] est différente de la relation [N AdjR]. Par exemple, dans la phrase "François a des jambes éléphantesques", l'adjectif "éléphantesques" n'établit pas une relation entre les jambes de François et la catégorie "éléphant", il leur attribue une qualité des individus de cette catégorie : être très gros, exemple extrait de Roché (2006).

Dubois et Dubois-Charlier (1999, p. 129) et Goes (1999, p. 251) citent certaines propriétés de ces adjectifs que nous résumons sous le titre de propriétés linguistiques et présentons dans la table 1 (P1 à P6). D'autres propriété que nous appelons "opérationnelles" et qui se sont basées sur les propriétés linguistiques sont présentées également dans la table 1. Les propriétés opérationnelles ne sont pas toujours exclusives aux adjectifs relationnels mais elles nous permettent de repérer des adjectifs automatiquement dans un corpus.

Les adjectifs relationnels sont dérivés par suffixation d'un nom. Les suffixes des adjectifs relationnels peuvent être : *-ien, -ois, -ique, etc.* (P7 dans la table 1). Toutefois, la détection automatique des noms de base dont les adjectifs relationnels sont dérivés ne se fait pas par une simple comparaison entre la base nominale et l'adjectif relationnel dé-suffixé à cause de l'allomorphie des bases ; "*l'addition d'un suffixe peut entraîner des modifications morphologiques de la base nominale, elles sont plus ou moins importantes selon la nature de N ou selon la nature du suffixe*" (Dubois et Dubois-Charlier, 1999, p. 135). Par exemple, ces modifications peuvent être : la modification phonique ou graphique de *N* (tropique/tropical), l'addition de voyelles ou de syllabes (nom/nominal), modification du radical à partir du latin (bête/bestial), etc. Par ailleurs, les adjectifs relationnels et les adjectifs déverbaux ont quelques suffixes en commun, qui sont : *-if, -aire, -eux, -oire, et -é*. La catégorie d'un adjectif ne peut donc pas être déterminée en ne s'appuyant que sur son suffixe.

2.2 Identification

La tâche d'identification des adjectifs relationnels dans un corpus n'est pas simple : d'une part la classe des adjectifs relationnels est floue, et d'une autre, il n'y a pas de règles véritablement sûres pour les identifier automatiquement (Goes, 1999; Maniez, 2005). De plus, les adjectifs relationnels dérivent, avec le temps, de façon régulière vers la qualification (Noailly, 1999, p. 24). Par exemple, certains adjectifs peuvent jouer un rôle relationnel ou qualificatif selon le contexte (ex. le système nerveux (*AdjR*) vs. François est nerveux (*AdjQ*)). Un adjectif peut donc avoir dans un terme deux interprétations, l'une relationnelle et l'autre qualificative, par exemple, "une chaise royale" : est-elle la chaise du roi ou une chaise luxueuse ? Si on identifie l'adjectif "royale" comme relationnel, et qu'on l'aligne avec le nom "roi" quand il s'agit d'une utilisation qualificative de cet adjectif : "chaise royale" sera paraphrasé par "chaise du roi" qui peut être traduit en anglais par "chair of the king". L'alignement d'un adjectif qualificatif avec un nom peut donc introduire de mauvaises traductions pour la méthode compositionnelle. Cependant, quand un adjectif peut avoir un emploi relationnel, l'alignement de cet adjectif avec son nom de base peut aider à la traduction des termes [N A] avec une haute précision (Morin et Daille, 2010).

Plusieurs travaux se sont intéressés à l'identification des adjectifs relationnels, nous présentons brièvement ci-dessous les travaux de Daille (1999) et Maniez (2005) qui se penchent sur l'extraction automatique ou semi-automatique des adjectifs relationnels à partir des corpus monolingues, ainsi que le travail de Cartoni (2008) sur les mots préfixés.

Daille (1999) exploite des règles de désuffixation-recodage (définies manuellement pour le français et l'anglais) pour relier un adjectif relationnel avec son nom de base (ex. la règle (-estière, -êt) peut relier "forestière" à "forêt"). Un adjectif *A* extrait à l'aide de ces règles, et qui doit apparaître avec un nom recteur *X* sous la forme [X A], sera considéré comme relationnel s'il peut être paraphrasé par un groupe [préposition substantif] sous la forme [X PREP DET ? N′] ; où *N′* est le nom dont *A* est dérivé

(voir P8 dans la table 1). La recherche des paraphrases est faite à partir du corpus. Cette méthode donne une précision de 99 %, mais un faible rappel dû au nombre limité de paraphrases dans le corpus.

Maniez (2005) examine deux approches pour identifier les adjectifs relationnels dans un corpus de spécialité en anglais : (a) il se penche sur l'hypothèse que dans un corpus spécialisé, la plupart des adjectifs sont relationnels. Ainsi, il exploite P1 et P4 (voir la table 1) afin de filtrer les adjectifs non-relationnels dans le corpus (b) tous les adjectifs en deuxième position extraits à partir du motif [ADJ1-ADJ2-N] sont sélectionnés en tant qu'adjectifs relationnels. Ce motif peut être adapté en français par [N-ADJ1-ADJ2], et nous ajoutons le critère suivant : si *ADJ*2 est relationnel, ADJ1 est également relationnel. La raison pour laquelle nous considérons que l'adjectif en première position est relationnel, c'est parce que l'adjectif relationnel suit immédiatement le nom (Pedreira, 2002), et qu'on détermine avant de qualifier (ex. un discours présidentiel intéressant), nous conclurons donc qu'un adjectif qualificatif ne peut pas précéder un adjectif relationnel, cette propriété est décrite sous P9 dans la table 1.

Cartoni (2008) travaille sur les mots préfixés qui ont la forme : [préfixe M] (ex. antitumoral). Il constate qu'avec certains préfixes (comme *post-*), si *M* est un adjectif, il s'agit d'un adjectif relationnel. Avec un autre groupe de préfixes (comme *anti-*), *M* est soit un adjectif relationnel, soit un adjectif déverbal (Cartoni, 2008, p. 255)(voir P11 et P12 dans la table 1).

Nous allons en premier développer dans la section suivante une méthode pour extraire une liste des adjectifs relationnels du corpus à l'aide des propriétés présentées.

Propriétés linguistiques	
P1	"*ils n'acceptent pas d'adverbe de degré*" (acide très nitrique, sauf cas particulier) (Dubois et Dubois-Charlier, 1999). Les adjectifs relationnels "*refusent la gradation en général, et "très" en particulier*" (Goes, 1999)
P2	"*ils ne peuvent pas être antéposés*" (le nitrique acide).
P3	"*ils ne sont pas susceptibles d'adverbialisation*" (nitriquement) "*ni de nominalisation*" (nitricité).
P4	"*ils ne s'emploient pas en fonction d'attribut*" (cet acide est nitrique, sauf cas particulier).
P5	"*la coordination d'un adjectif relationnel avec un adjectif qualificatif est impossible*".
P6	"*ils ne forment généralement pas de séries antonymes*".
Propriétés opérationnelles	
P7	les suffixes des adjectifs relationnels : *-ique, -aire, -eux, -ier, -ien, -ois, -ain, -al, -el, -estre, -il, -in, -esque, -é, -if*.
P8	il existe des paraphrases dans un corpus monolingue de la forme [X PREP DET ? N′] : : [X AdjR] (ex. cancer du poumon : : cancer pulmonaire) ; où *X* est un nom, N′ est le nom de base de *AdjR*. (*PREP* signifie préposition et *DET* signifie déterminant, *?* signifie que *DET* peut apparaître une ou zéro fois)
P9	dans les syntagmes de la forme [N Adj1 Adj2] (ex. rupture capsulaire ganglionnaire), si *Adj*2 est un adjectif relationnel, *Adj*1 est relationnel également.
P10	dans les syntagmes de la forme [N Adj1 et/ou Adj2] (ex. facteurs environnementaux ou génétiques), si *Adj*2 est un adjectif relationnel, *Adj*1 est relationnel également.
P11	ils peuvent être préfixés par les préfixes : *post-, trans-, uni-, tri-, anti-, tri-, pré-*.
P12	ils peuvent être préfixés par des racines gréco-latines : *micro-, séro-, radio-, etc*.

TABLE 1 – Propriétés linguistiques et opérationnelles des adjectifs relationnels

3 Extraction des adjectifs relationnels du corpus

La reconnaissance automatique des *AdjR* en corpus pose un certain nombre de problèmes comme nous l'avons vu en section 2 : (a) ambiguïté des suffixes ; (b) ambiguïté de la classe relationnel/qualificatif ; (c) indice de relation exprimé par les propriétés non-présentes en corpus. Dans cette section, nous

développons une approche pour extraire des adjectifs relationnels automatiquement du corpus.

3.1 Approche

Afin d'extraire des adjectifs relationnels du corpus, nous exploitons quelques propriétés linguistiques et opérationnelles présentées dans la table 1. Nous partons de l'hypothèse que les racines gréco-latines et certains préfixes français préfixent des adjectifs non-qualificatifs pour extraire une liste d'adjectifs initiale (en utilisant les propriétés P11 et P12). Nous nous basons sur cette liste afin de l'étendre en utilisant d'autres propriétés (P9 et P10). La méthode d'identification automatique des adjectifs relationnels du corpus que nous proposons est présentée dans l'algorithme 1 (nous faisons référence dans cet algorithme aux propriétés listées dans la table 1). L'ensemble des listes que nous extrayons sera utilisé par l'approche d'alignement adjectif-nom présentée en section 4.

3.1.1 Remarques

1. Il y a des racines (ex. "bio-") qui peuvent préfixer des adjectifs déverbaux (ex. biodégradable). Cependant, dans le cas des adjectifs préfixés par ces racines et qui se terminent par un suffixe qui ne peut pas être déverbal (ex. -ique dans "biochimique"), nous considérons que ces adjectifs sont relationnels.
2. Afin de trouver les adverbes construits à partir d'un adjectif dans le corpus : (a) nous ajoutons le suffixe adverbial "ment" (et d'autres adaptations du suffixe) à l'adjectif (b) nous cherchons ces adverbes construits dans le corpus.
3. L'extraction des adjectifs relationnels par le biais de la propriété P9 est plus fiable que l'extraction de ces adjectifs par P10. En effet, P10 peut introduire du bruit quand il s'agit d'une utilisation qualificative d'un adjectif. Pour cette raison, nous choisissons de ne l'appliquer que sur les adjectifs relationnels qui sont trouvés à l'aide de P11 et P12 (qui ont en effet peu de chances d'avoir un emploi qualificatif).
4. Bien qu'on puisse aussi extraire des adjectifs déverbaux par cette méthode, on peut les relier la plupart du temps à des substantifs (ex. végétatif ; Verbe : végéter ; Nom : végétation).

3.2 Identification des racines gréco-latines

Nous visons à extraire une liste de racines $L_{racines}$ automatiquement du corpus, ces racines sont utilisées par l'algorithme d'identification automatique des adjectifs relationnels présenté dans l'algorithme 1. Nous supposons que les racines gréco-latines préfixent les bases adjectivales non-qualificatives. Certaines racines ne peuvent préfixer que des adjectifs relationnels, alors que d'autres peuvent préfixer des adjectifs relationnels ou déverbaux. Nous présentons la méthode que nous développons pour extraire des racines dans l'algorithme 2.

4 Alignement d'un adjectif relationnel avec un nom

Nous supposons qu'un adjectif relationnel partage des caractères dans le même ordre avec son nom de base. Afin de trouver le nom *N* dont un adjectif *A* est dérivé, on peut comparer cet adjectif avec tous les noms dans un dictionnaire (et qui existent dans un corpus source) en donnant des scores entre un nom et un adjectif par des mesures de similarité. Le nom *N* qui a la similarité la plus élevée avec *A* sera retenu. De nombreux algorithmes existants peuvent mesurer la similarité ou la distance entre deux chaînes. En effet, une mesure intéressante pour notre tâche préservera l'ordre linéaire des lettres lors de la comparaison de deux chaînes. Nous présentons ci-dessous deux mesures qui préservent l'ordre, ces mesures seront utilisées dans la suite :

1. Lcs : cette mesure consiste à trouver la sous-séquence la plus longue "longest common subsequence[3]" entre deux chaînes, l'ordre des lettres est donc préservé. Par exemple, Lcs(forestier,

3. subsequence : les lettres de la sous-séquence sont dans le même ordre que dans la chaine complète, substring : les lettres de la sous-chaîne sont consécutives et dans le même ordre que dans la chaine complète.

Données : $L_{prefixes}$ (préfixes français qui n'acceptent qu'une base adjectivale relationnelle ou déverbale), $L_{suffRel}$ (suffixes relationnels), $L_{racines}$ (racines gréco-latines extraites automatiquement, voir 3.2), $C_{cds_{fr}}$ (corpus français), $L_{[NA]_{fr}}$ (termes [N A] extraits du corpus français);
Résultat : $Liste_{AdjR}$-1 (contient les adjectifs qui commencent par une racine dans $L_{racines}$ ou un préfixe dans $L_{prefixes}$), $Liste_{AdjR}$-2 (contient les adjectifs qui peuvent être préfixés par des racines dans $L_{racines}$ ou des préfixes dans $L_{prefixes}$), $Liste_{AdjR}$-3 (contient les adjectifs extraits à l'aide de P9), $Liste_{AdjR}$-4 (contient les adjectifs relationnels extraits à l'aide de P10);
début
  pour *chaque A qui apparait dans au moins un terme [N A]* $\in L_{[NA]_{fr}}$, *et qui se termine par un suffixe* $\in L_{suffRel}$ *(ex. tumoral)* **faire**
    si *il existe un autre A″ (ex. hématotumoral) dans* $C_{cds_{fr}}$ *qui a la forme [racine A] ou [préfixe A] (ex. racine=hémato, A=tumoral) (où préfixe* $\in L_{prefixes}$, *racine* $\in L_{racines}$*)* **alors**
      si *le "préfixe" ou la "racine" n'accepte que des bases relationnelles* **alors**
        Ajouter *A″* à $Liste_{AdjR}$-1 et *A* à $Liste_{AdjR}$-2;
      sinon
        si *le suffixe de A est un suffixe non-commun entre les adjectifs relationnels et les adjectifs déverbaux (ex. le suffixe "ique", voir 1 dans 3.1.1)* **alors**
          Ajouter *A″* à $Liste_{AdjR}$-1 à et *A* à $Liste_{AdjR}$-2;
  $temp_{AdjR} \longleftarrow \{ Liste_{AdjR}\text{-}1 \cup Liste_{AdjR}\text{-}2 \}$;
  répéter
    pour *chaque adjectif Ad jR dans* $temp_{AdjR}$ *qui vérifie P1 (avec "très")* **faire**
      $tempList_{A''} \longleftarrow$ Trouver tous les adjectifs qui précèdent *Ad jR* immédiatement dans les motifs de la forme : [N *A″* AdjR] (ex. profil protéique tumoral, voir P9);
      pour *chaque* $A'' \in tempList_{A''}$ **faire**
        si *A″ respecte les propriétés P1 (avec très) et P3 (voir 2 dans 3.1.1)* **alors**
          Ajouter *A″* à $Liste_{AdjR}$-3;
          $temp_{AdjR} \longleftarrow temp_{AdjR} \cup Liste_{AdjR}$-3;
  jusqu'à *Pas de nouveaux adjectifs ajoutés à* $Liste_{AdjR}$-3;
  pour *chaque adjectif Ad jR dans* $Liste_{AdjR}$-1 **faire**
    $tempList_{A''} \longleftarrow$ Trouver tous les adjectifs qui sont en coordination avec *Ad jR* (ex. mammaire et tumoral, voir P10);
    pour *chaque* $A'' \in tempList_{A''}$ **faire**
      si *A″ respecte les propriétés P1 (avec très) et P3* **alors**
        Ajouter *A″* à $Liste_{AdjR}$-4;

Algorithme 1: Identification des adjectifs relationnels dans le corpus

forêt)=fort. On peut normaliser cette mesure comme suit (Ketkar et Youngblood, 2010) : $Lcs_{normalise}(A,B) = |Lcs(A,B)|^2/(a \times b) \in [0,1]$, où a est la longueur de A et b la longueur de B. Plus ce score est élevé, plus les chaînes sont similaires.
Exemple : $Lcs_{normalise}$(forestier, forêt)=16/45=0,35.

2. Levenshtein : cette distance est définie comme le nombre minimum de modifications nécessaires pour transformer une chaîne en une autre. Les opérations autorisées sont l'insertion, la suppression et la substitution d'un seul caractère. Le coût de chaque opération est égal à 1. Par exemple, Levenshtein(forestier, forêt)=5. On peut normaliser cette distance ($\in [0,1]$) si on la divise par la chaîne la plus longue. Moins ce score est élevé, plus les chaines sont similaires.
Exemple : $Lev_{normalise}$(forestier, forêt)=5/9=0,55.

Données : $C_{cds_{fr}}$ (corpus français), $L_{[NA]_{fr}}$ (termes français de la forme [N A]), $L_{suffRel}$ (suffixes relationnels) ;
Résultat : $L_{racines}$;
début
  pour *chaque adjectif A dans* $C_{cds_{fr}}$ *qui compose un terme [N A] dans* $L_{[NA]_{fr}}$ *(ex. barrière tumoral)* **faire**
    si *il se trouve un autre adjectif A′ dans le corpus (ex. hématotumoral), où A′ peut s'écrire de la forme suivante : [élément A′] (ex. hémato tumoral), et si cet élément se termine par "o", et s'il n'est pas l'un des préfixes français qui se terminent par "o" : hypo-, rétro- ou pro-* **alors**
      Ajouter "élément" (ex. hémato) à $L_{racines}$;
      si "*élément*" *préfixe*
      au moins un adjectif Adj dans $C_{cds_{fr}}$ *où Adj se termine par un suffixe* $\notin L_{suffRel}$ **alors**
        "élément" est une racine qui peut préfixer les adjectifs déverbaux (ex. bio-);
      sinon
        "élément" est une racine qui ne préfixe que les adjectifs relationnels (ex. micro-);

Algorithme 2: Identification des racines gréco-latines

On peut prendre (1-$Lev_{normalise}$) pour mesurer la similarité entre deux chaînes.

Les mesures de similarité sont souvent utilisées dans la tâche d'identification des cognats entre deux langues (mots similaires orthographiquement et qui ont un sens similaire, ex. FR activiste / EN activist). Par exemple, Hauer et Kondrak (2011) et Frunza et Inkpen (2010) utilisent ou combinent plusieurs mesures de similarité telles que Levenshtein, Lcs, Soundex, Longest prefix, etc. Les scores obtenus entre chaque couple de mots seront ensuite utilisés comme traits par un algorithme d'apprentissage qui les classifient comme cognats ou non. Afin d'identifier les mots qui sont des faux cognats, Frunza et Inkpen (2010) utilisent un corpus parallèle qui sert à désambiguïser les sens des mots. Les mesures de similarité ont été également utilisées par Cartoni (2009) pour relier un adjectif relationnel avec son nom de base. Cependant, Cartoni (2009) ne traite pas les adjectifs relationnels construits à partir des formes supplétives des noms. En outre, il exige qu'un adjectif et un nom aient une similarité de lettres très importante afin de les relier automatiquement.

Dans la suite de cette section, nous développons des approches pour aligner un adjectif avec un nom. Nous nous appuyons d'abord sur la similarité de lettres entre un adjectif et un nom. Nous utilisons ensuite la propriété P8 (voir la table 1) en nous inspirant du travail de Daille (2000) afin que les scores obtenus par les mesures de similarité soient plus répresentatifs. Enfin, nous utilisons des racines gréco-latines pour relier les adjectifs relationnels supplétifs avec des noms.

4.1 Alignement adjectif-nom par mesures de similarité de lettres

D'abord, nous essayons de relier un adjectif avec un nom en n'utilisant que des mesures de similarité. Nous combinons les deux similarités : $similarity_{Lcs}(A,B)$ et $similarity_{Lev}(A,B)$ (voir ci-dessous), en prenant leur moyenne géométrique afin d'avoir un seul score $\in [0,1]$ entre un adjectif et un nom :

$$similarity_{lettres}(A,B) = (similarity_{Lcs}(A,B) + similarity_{Lev}(A,B))/2 \quad (1)$$

$$similarity_{Lcs}(A,B) = |Lcs(A,B)|^2/(a \times b) \quad (2)$$

$$similarity_{Lev}(A,B) = \begin{cases} 1-(Levenshtein(A,B)/a) & \text{si } a >= b \\ 1-(Levenshtein(A,B)/b) & \text{autrement} \end{cases} \quad (3)$$

Où $similarity_{Lcs}(A,B)$ et $similarity_{Lev}(A,B) \in [0,1]$, a et b sont les longueurs des chaînes A et B respectivement. Dans le calcul du score de Levenshtein, chaque opération a un coût égal à 1.

Cependant, nous donnons une pénalité moins élevée à la substitution de deux lettres qui sont proches phonétiquement. Par exemple, on fixe la pénalité de la substitution de "f" par "v" et "é" par "è" à 0,5. Nous nous inspirons de Dubois et Dubois-Charlier (1999) pour définir ces substitutions, puisque des adaptations générales de la langue française y sont définies.

Si un adjectif peut avoir un emploi nominal (considéré comme un substantif dans le dictionnaire), nous l'alignons avec lui-même (ex. clinique, esthétique, etc). De plus, nous supposons qu'un adjectif relationnel commence par la même lettre que son nom de base. En effet, nous avons trouvé en examinant une liste de 200 adjectifs que cette hypothèse est vraie dans 97 % des cas.

Nous considérons qu'un adjectif est relié à un nom si le score entre les deux est supérieur ou égal à un certain seuil (on supprime le suffixe relationnel de l'adjectif lors de la comparaison). Plus on augmente le seuil de similarité plus le rappel est faible. La mesure Lcs favorise les noms les plus longs quand on compare un adjectif avec les noms du dictionnaire. Par exemple, selon Lcs, le nom "notion" est plus proche de "nominal" que de "nom" : |Lcs(nomin, notion)|=4, |Lcs(nomin, nom)|=3. Alors que les deux chaînes ont le même score avec "nominal" selon Levenshtein : Levenshtein(nomin, notion)=2, Levenshtein(nomin, nom)=2). De plus, si on exige une similarité très importante entre un adjectif et un nom, on pourra perdre de nombreux alignements corrects (ex. axillaire/aisselle, germe/germinal, etc). Pour cela, il faut choisir un seuil de similarité qui ne soit pas très important afin de permettre à d'autres méthodes de filtrage de mieux classifier les alignements obtenus par les mesures de similarité.

4.1.1 Alignement adjectif-nom par mesure de similarité contextuelle

Dans de nombreux cas, la similarité de lettres seule ne suffit pas pour trouver le nom avec lequel un adjectif est relié. Par exemple, comment peut-on dire que l'adjectif "sérique" est dérivé de "sérum" et non pas de "série" ? Nous essayons donc de modifier le score entre un adjectif relationnel et un nom dans un dictionnaire si la similarité entre le nom et l'adjectif est supérieure à un certain seuil, en cherchant des paraphrases monolingues dans lesquelles les deux mots apparaissent. Pour un adjectif A et un nom N, nous cherchons une paraphrase dans le corpus de la forme [X A : : X PREP DET ? N] [4], où X est un nom tête. Comme par exemple, cancer pulmonaire : : cancer du poumon. Afin de calculer un score entre A et N, nous représentons chacun par un vecteur où les attributs sont les noms têtes qui apparaissent avec A ou N (voir figures 2). Le score d'un attribut dans le vecteur d'un nom N est calculé à l'aide de la mesure d'association IM entre N et le nom tête X :

$$IM(N,X) = \log_2 \frac{a}{(a+b)(a+c)} \tag{4}$$

– a est le nombre d'occurrences de N et X ensemble
– b est le nombre d'occurrences de N avec tous les autres nom têtes $\neq X$
– c est le nombre d'occurrences de X avec tous les autres noms $\neq N$

Le score de chaque attribut dans un vecteur d'adjectif A est calculé de la même manière :

$$IM'(A,X) = \log_2 \frac{a'}{(a'+b')(a'+c')} \tag{5}$$

– a' est le nombre d'occurrences de A et X ensemble
– b' est le nombre d'occurrences de A avec tous les autres nom têtes $\neq X$
– c' est le nombre d'occurrences de X avec tous les autres adjectifs $\neq A$

Ensuite, nous calculons un score entre les deux vecteurs (nom et adjectif) en utilisant le cosinus.

$$similarity_{paraphrases}(A,N) = \cos(A,N) = \frac{\sum_{i=1}^{n} IM.IM'}{\sum_{i=1}^{n} IM^2 . \sum_{i=1}^{n} IM'^2} \tag{6}$$

4. On peut aussi inclure d'autres variantes, comme par exemple les formes [X A_1 A] (ex. région ganglionnaire axillaire) et [X A_1 PREP DET N] (ex. balayage lent de l'aisselle).

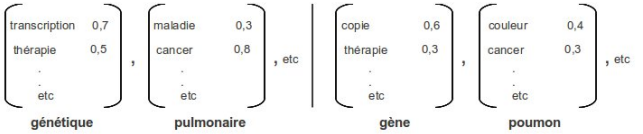

FIGURE 2 – Vecteurs des adjectifs relationnels et des noms

Où n est le nombre de noms têtes communs entre A et N.

4.1.2 Combinaisons des mesures de similarité de lettres et de paraphrases

Pour un adjectif A et un nom N, nous calculons leur score final en combinant leur similarité de lettres selon la formule 1 et la similarité contextuelle selon la formule 6, comme suit :

$$score(A,N) = \alpha * similarity_{lettres}(A,N) + \beta * similarity_{paraphrases}(A,N) \quad (7)$$

En effet, le choix des valeurs de α et β dépend des seuils minimaux choisis pour la similarité de lettres. Par exemple, si on permet une différence de lettres importante, il faut donc choisir $\alpha > \beta$.

Un adjectif A sera relié au nom avec lequel il a le score le plus élevé.

4.1.3 Alignement adjectif-nom en utilisant des racines supplétives

Certains adjectifs contiennent des racines supplétives, et dans certains cas il n'est pas possible de les relier à leurs noms en s'appuyant sur la similarité de lettres quand la modification du nom de base par la racine supplétive est importante (ex. médullaire/moelle).

Nous considérons qu'un adjectif A qui commence par une séquence de lettres identique à une racine supplétive est relationnel s'il remplit une des conditions suivantes :

– sa forme peut être identifiée comme étant : [racine suffixe] (où *racine* est une des racines supplétives dans une liste des racines alignées avec des noms ($L_{racines-noms}$) (ex. pulmon/poumon, médull/moelle, etc), et *suffixe* est un des suffixes relationnels dans une liste de suffixes relationnels). Par exemple, l'adjectif "pulmonaire" peut être décomposé en "pulmon" (une racine) et "aire" (un suffixe).
– il construit avec le nom N associé à la racine supplétive (selon $L_{racines-noms}$) au moins une paraphrase de la forme [X A : : X PREP DET ? N], où X est un nom tête. Par exemple, biopsie de moelle : : biopsie médullaire.

4.1.4 Combinaison des méthodes d'alignement

Nous combinons l'approche décrite en section 4.1.3 et celle présentée dans la section 4.1.2 comme suit : pour un adjectif A, nous vérifions s'il peut être relié avec un nom à l'aide des racines supplétives, sinon, on applique la formule 7 entre A et chaque nom dans le dictionnaire qui existe dans le corpus.

De plus, si A (ex. oncogénique) commence par un préfixe ou une ou plusieurs racines supplétives, et s'il peut s'écrire de la forme : [(racine | préfixe)+ A′] [5](ex. oncogénique), où A' (ex. génique) est un adjectif dans le corpus : on relie A' à un nom N' (ex. gène), ensuite, on cherche le nom n = [(racine | préfixe)+ N′] (ex. oncogène) dans le corpus, si ceci est trouvé, le nom de base avec lequel A est aligné sera n.

Nous supposons aussi que deux adjectifs qui partagent la même base (ex. sérique/séreux (sér), soigneux/soigné (soign), cellulaire/celluleux (cellul), etc.), doivent être alignés avec le même nom de base, sinon on considère que les alignements sont mauvais et on les supprime de la liste.

5. | signifie "ou", + signifie "1 à plusieurs"

5 Traduction des termes [N AdjR] en utilisant des alignements adjectif-nom

Nous utilisons les alignements adjectif-nom que nous obtenons par l'approche d'alignement adjectif-nom dans la tâche de traduction compositionnelle de termes complexes de la forme [X AdjR].

5.1 Approche

Pour chaque t_c=[N AdjR], nous remplacons $AdjR$ par le nom N' avec lequel il a été aligné. Ce remplacement donne un nouveau terme complexe $t_c{}'$=[N PREP DET ? N'], nous supposons que sa traduction est équivalente à celle de t_c. Nous suivons l'algorithme présenté sous algorithme 3.

Données : $L_{[NA]_{fr}}$ (termes français de la forme [N A]), $L_{[AN]_{en}}$ (termes anglais de la forme [A N]), $L_{[NN]_{en}}$ (termes anglais de la forme [N N]), $L_{alignements}$ (alignements adjectif-nom), $Dico_{fr-en}$ (dictionnaire bilingue français-anglais) ;
Résultat : Liste de traductions;
début
  pour *chaque expression de forme* $[N\,A] \in L_{[NA]_{fr}}$ *(ex. FR concentration plasmatique)* **faire**
    $N_{en} \longleftarrow$ traduire N par $Dico_{fr-en}$ (ex. $N_{en} \longleftarrow$ EN concentration);
    $A_{en} \longleftarrow$ traduire A par $Dico_{fr-en}$ (ex. $A_{en} \longleftarrow$ EN plasmatic);
    si $[A_{en}\ N_{en}]$ *(ex. plasmatic concentration)) existe dans* $L_{[AN]_{en}}$ **alors**
      [A_{en} N_{en}] est la traduction de [N A];
    sinon
      Prendre le nom N'
      (ex. FR plasma) avec lequel l'adjectif A (ex. FR plasmatique) est aligné de $L_{alignements}$;
      $N'_{en} \longleftarrow$ traduire N' par $Dico_{fr-en}$ (ex. $N_{en} \longleftarrow$ EN plasma);
      si $[N'_{en}\ N_{en}]$ *(ex. EN plasma concentration) existe dans* $L_{[NN]_{en}}$ **alors**
        [N'_{en} N_{en}] est la traduction de [N A];

Algorithme 3: Traduction des termes en utilisant les alignements adjectif-nom

6 Evaluation

Dans cette section, nous évaluons les approches que nous proposons pour (a) extraire des adjectifs relationnels (voir section 3), (b) aligner un adjectif extrait avec son nom de base (voir section 4), (c) traduire un adjectif relationnel dans les termes [N AdjR] en le remplaçant par le nom avec lequel il est aligné (voir section 5).

6.1 Ressources

Nous disposons des ressources suivantes :
– corpus comparable français-anglais dans le domaine du cancer du sein ($C_{cds_{fr}}$ et $C_{cds_{en}}$). Nous utilisons l'outil d'extraction et d'alignement des termes à partir des corpus Term Suite [6] (Rocheteau et Daille, 2011), afin d'étiqueter le corpus et d'extraire des termes. $C_{cds_{fr}}$ contient 14 680 mots distincts, alors que $C_{cds_{en}}$ contient 8 492 mots distincts. Nous extrayons des phrases selon des motifs, comme suit :
  – 12 991 phrases françaises ($L_{[NA]_{fr}}$) extraites par [N A], et 11 941 phrases anglaises ($L_{[AN]_{en}}$) extraites par [A N].
  – 12 954 phrases françaises ($L_{[NN]_{fr}}$) extraites par [N DET ? PREP N], et 10 069 phrases anglaises ($L_{[NN]_{en}}$) extraites par [N N].

6. http ://code.google.com/p/ttc-project/

- liste de préfixes $L_{prefixes}$ en français qui n'acceptent qu'une base adjectivale relationnelle ou déverbale, cette liste a été établie par (Cartoni, 2008).
- liste de 15 suffixes relationnels en français $L_{suffRel}$ (voir P7 dans la table 1).
- deux listes d'adjectifs français extraites automatiquement du corpus :
  - $LAdjR_{Classes}$: cette liste comprend 361 adjectifs extraits automatiquement du corpus. Elle correspond à l'ensemble des listes extraites du $C_{cds_{fr}}$ en suivant l'algorithme 1.
  - $LAdjR_{Base}$: contient tous les adjectifs extraits à partir de la propriété P7 (c'est-à-dire à partir des suffixes relationnels) et qui composent au moins un terme [N A]$\in L_{[NA]_{fr}}$. Cette liste contient 1 346 adjectifs, elle est considérée comme la liste de base et les résultats de l'alignement adjectif-nom sur cette liste seront comparés avec ceux obtenus sur la liste $LAdjR_{Classes}$.
- dictionnaire bilingue français-anglais ($Dico_{fr-an}$) de 145 542 entrées de mots simples.
- liste de 66 racines supplétives françaises ($L_{racines-noms}$) alignées avec des noms communs (ex. hépat / fois, pulmon / poumon, ... etc) (Cottez, 1982).
- liste de 100 racines $L_{racines}$ extraite automatiquement du $C_{cds_{fr}}$ en appliquant l'algorithme 2 présenté en section 3.2.

6.2 Résultats de l'extraction automatique des adjectifs relationnels

En appliquant l'algorithme d'extraction des adjectifs présenté en section 3 sur $C_{cds_{fr}}$, nous obtenons quatre listes d'adjectifs. Un adjectif extrait appartient à une ou plusieurs classes d'adjectifs : qualificative, relationnelle, composée. Par exemple, l'adjectif "sérologique" est composé et relationnel, car il peut être relié à "sérologie" et il se compose de deux éléments : "séro" et "logique". Les adjectifs de la classe "composée" ont des emplois non-qualificatifs, mais dans certains cas, on ne peut pas les relier avec un seul substantif, mais avec un syntagme, par exemple, "unilatéral" (un seul côté) ou "infraclinique" ("un trouble ou d'une maladie qui ne provoque pas de manifestation décelable à l'examen") n'ont pas été formés par dérivation d'un nom mais par préfixation.

Les listes extraites sont présentées dans la table 2. Nous appelons l'ensemble de ces listes $LAdjR_{Classes}$ qui comprend donc 361 adjectifs. Les adjectifs ont été classés manuellement et à l'aide du système Dérif (Namer, 2003). Nous remarquons que 198 adjectifs dans $LAdjR_{Classes}$ peuvent être classifiés comme relationnels, et qu'il y a beaucoup d'adjectifs composés qui ne sont ni relationnels ni qualificatifs.

La liste $LAdjR_{Classes}$ contient plus de 54 % d'adjectifs relationnels et la liste $Liste_{AdjR}$-2 se compose de 93 % d'adjectifs relationnels. Pour avoir une idée du rappel, nous utilisons Dérif pour aligner les adjectifs de la liste $LAdjR_{Base}$ avec des noms. Dérif est capable d'aligner 554 adjectifs avec des noms par la relation "en rapport avec". Nous appelons cette liste par $Liste_{Derif}$. Nous trouvons que la liste $LAdjR_{Classes}$ couvre 141 adjectifs de $Liste_{Derif}$. Cependant, 57 des adjectifs que nous avons classifiés comme relationnels dans $LAdjR_{Classes}$ n'ont pas pu être alignés par Dérif. De plus, il existe des adjectifs dénominaux mais non relationnels dans $Liste_{Derif}$ (ex. original/origine, critique/crise, etc.).

liste	nbr. d'adjectifs	nbr. de classe qualificative	nbr. de classe relationelle	nbr. classe composée
$Liste_{AdjR}$-1	154	0	28	153
$Liste_{AdjR}$-2	103	8	96	19
$Liste_{AdjR}$-3	47	3	34	18
$Liste_{AdjR}$-4	57	6	40	27
Total	361	17	198	217

TABLE 2 – Les classes des adjectifs dans les listes extraites

Les listes d'adjectifs extraites seront utilisées par la méthode de l'alignement d'un adjectif relationnel avec un nom.

6.3 Résultats de l'alignement adjectif-nom sur les listes d'adjectifs

Nous appliquons la méthode d'alignement que nous avons proposée en section 4.1.4 sur les listes des adjectifs extraits automatiquement ($LAdjR_{Classes}$ et $LAdjR_{Base}$). Nous fixons empiriquement les poids des deux similarités dans l'équation 7 : α=0,70 et β=0,30. Un adjectif et un nom doivent avoir une similarité minimale de $similarity_{Lev}$ à 0,6 et une similarité minimale de $similarity_{Lcs}$ à 0,7 (les deux similarités qui composent $similarity_{lettres}$ dans l'équation 7).

Ainsi, 157 adjectifs de la liste $LAdjR_{Classes}$ (parmi 361) ont été alignés avec une précision de 89,8 %. De la liste $LAdjR_{Base}$, 582 adjectifs (parmi 1 346) ont été alignés avec une précision de 84,53 %. Nous avons évalué les alignements manuellement et à l'aide de l'outil Dérif Namer (2003). Nous considérons qu'un alignement est correct si l'adjectif a été aligné avec lui-même ou avec son nom de base. En effet, $LAdjR_{Base}$ contient plus des adjectifs non-relationnels et du bruit (des mots non-français) que $LAdjR_{Classes}$, ce qui explique le taux plus élevé des mauvais alignements. De plus, nous exigeons que le nom de base d'un adjectif soit présent dans le corpus, alors que ce n'est pas toujours le cas. Le rappel est le nombre d'adjectifs alignés divisé par le nombre d'adjectifs dans la liste. Cependant, il faut noter qu'il y a de nombreux adjectifs dans $LAdjR_{Base}$ et $LAdjR_{Classes}$ qui ne peuvent pas être reliés à des noms. Par exemple, les adjectifs composés sont parfois reliés à des phrases comme on l'avait déjà mentionné dans la section 6.2. Nous résumons les résultats de l'alignement dans la table 3.

liste	nbr. d'alignements adj-nom	précision	rappel
$LAdjR_{Classes}$	157	89,8 %	43,49 %
$LAdjR_{Bases}$	582	84,53%	43,23 %

TABLE 3 – Résultats des méthodes d'alignement adjectif-nom sur $LAdjR_{Classes}$ et $LAdjR_{Base}$

Nous présentons les résultats de la traduction des termes [N AdjR], en utilisant les alignements adjectif-nom obtenus, dans la section suivante.

6.4 Résultats de la traduction des termes [N AdjR]

La méthode compositionnelle qui consiste à traduire des termes français de la forme [N A] en termes anglais de la forme [A N] nous a permis de traduire 2 039 termes dont les adjectifs sont issus de la liste $LAdjR_{Base}$, et 574 termes dont les adjectifs sont issus de la liste $LAdjR_{Classes}$. Cette méthode a donné une précision de 79,5 % sur une liste de 200 termes traduits qui a été examinée manuellement. Nous essayons maintenant de traduire les termes français [N A] non-traduits par la méthode précédente en passant par les noms de base des adjectifs relationnels.

Nous suivons l'algorithme 3 afin d'évaluer l'impact des alignements adjectif-nom sur la traduction des termes [N A], voir la table 4. Nous utilisons les 157 alignements adjectif-nom obtenus de la liste $LAdjR_{Classes}$ et nous trouvons que 42 alignements adjectif-nom de cette liste ont aidé à traduire 172 termes [N A] distincts avec une précision de 91,86 %. En appliquant l'algorithme 3 sur les 582 alignements adjectif-nom obtenus de $LAdjR_{Base}$, nous trouvons que 92 de ces alignements ont participé à traduire 250 termes distincts avec une précision de 86 %. Les traductions ont été vérifiées à l'aide du dictionnaire rédactionnel Linguee [7] et la banque de données Termium [8]. La précision des traductions est égale au nombre de termes distincts qui ont au moins une traduction correcte parmi les 5 premières traductions proposées divisé par le nombre de termes distincts qui ont été traduits. Les traductions proposées ont été classées par leurs fréquences dans le corpus cible.

Les alignements des adjectifs dénominaux qui ont des emplois qualificatifs (ex. originale/origine,

7. http ://www.linguee.fr/
8. http ://www.termiumplus.gc.ca/tpv2alpha/alpha-fra.html ?lang=fra

formel/forme) avec des noms ont donné des mauvaises traductions. Des adjectifs déverbaux qui peuvent être reliés à un nom, ont donné des bonnes et/ou des mauvaises traductions. Par exemple, l'adjectif "étudié" est dérivé du verbe "étudier", il a été relié avec le nom "étude" par la méthode d'alignement adjectif-nom. Cet alignement a donné de bonnes traductions (ex. "population étudiée" a été traduit par "study population"), ainsi que de mauvaises traductions (ex. "cellule étudiée" a été traduit par "study unit").

Parfois on trouve des mauvaises traductions malgré l'utilisation d'un alignement correct d'un adjectif relationnel avec un nom. Ces mauvaises traductions sont plutôt obtenues à cause des problèmes liés à la méthode compositionnelle et au corpus comparable. Par exemple, l'adjectif relationnel "génétique" a été relié avec "gène", cet alignement a participé à la traduction de "mutation génétique" par "gene transfer" ("gène" traduit par "gene") tandis que la bonne traduction est "gene mutation". Ainsi, la mauvaise traduction n'a pas été obtenue à cause de l'alignement de "génétique" avec "gène", mais parce que soit "mutation" n'a pas été traduit par "mutation" dans le dictionnaire bilingue, soit "gene mutation" n'existe pas dans le corpus anglais.

liste	nbr. d'alignements adj-nom	nbr. de termes [N A] traduits	précision
$\text{LAdjR}_{Classes}$	157	172	91,86 %
LAdjR_{Bases}	582	250	86,00 %

TABLE 4 – Résultats de la traduction en utilisant les alignements adjectif-nom

7 Discussion et conclusion

Dans cet article, nous nous sommes intéressés à l'identification des adjectif relationnels et à l'alignement de ces adjectifs avec leurs noms de base. Nous avons également essayé de traduire des termes qui se composent d'un nom et d'un adjectif relationnel [N AdjR] en remplaçant $AdjR$ par son nom de base.

Nous avons développé une méthode qui exploite plusieurs propriétés des adjectifs relationnels pour les identifier en se basant sur un corpus monolingue. Nous avons extrait par cette méthode une liste d'adjectifs $LAdjR_{Classes}$. Une autre liste d'adjectifs $LAdjR_{Base}$ a été extraite en utilisant une liste de suffixes relationnels. Nous avons trouvé que la liste $LAdjR_{Classes}$ contient très peu d'adjectifs qualificatifs et moins de bruit que la liste $LAdjR_{Base}$.

Ensuite, nous avons développé une méthode afin d'aligner les adjectifs relationnels extraits avec leurs noms de base à partir d'un corpus monolingue. Nous nous sommes appuyés sur la similarité de lettres, la similarité contextuelle et des racines gréco-latines afin de relier un adjectif à un nom. Nous avons appliqué la méthode d'alignement sur les deux listes $LAdjR_{Base}$ et $LAdjR_{Classes}$, et nous avons acquis des couples d'adjectif-nom avec une précision supérieure à 84 %.

Enfin, nous avons exploité les alignements adjectif-nom obtenus pour traduire compositionnellement des termes de la forme [N AdjR]. La précision des alignements adjectif-nom obtenus à partir de $LAdjR_{Classes}$, ainsi que la traduction des termes [N AdjR] obtenus en utilisant ces alignements ont été plus élevées que celles des alignements et des traductions obtenues avec $LAdjR_{Base}$. Par contre, nous obtenons plus d'alignements avec $LAdjR_{Base}$ et donc plus de traductions par rapport à l'utilisation de $LAdjR_{Classes}$. Il semble donc que la méthode d'alignement adjectif-nom sur une liste d'adjectifs purement relationnels peut donner des alignements avec une haute précision et ainsi une haute précision pour la traduction compositionnelle des termes [N AdjR]. Les mauvais alignements adjectif-nom n'ont pas beaucoup influencé la précision des traductions de ces termes qui est de 86 % en utilisant $LAdjR_{Base}$. La traduction compositionnelle permet donc de filtrer les mauvais alignements adjectif-nom.

Dans ce travail, nous nous sommes concentrés sur la traduction des termes [N AdjR] pour le couple de langues français-anglais. Le principe de traduction par paraphrase de ces termes pour d'autres couples de langues devra être étudié pour en démontrer la généricité.

Remerciements

Ce travail a bénéficié de l'aide du septième programme cadre de la Commission européenne (FP7/2007-2013) (Grant Agreement no 248005).

Références

BALDWIN, T. et TANAKA, T. (2004). Translation by machine of complex nominals : Getting it right. *In ACL Workshop on Multiword Expressions : Integrating Processing*, pages 24–31.

BOWKER, L. et PEARSON, J. (2002). *Working with specialized language : a practical guide to using corpora*. London, Routledge.

CARTONI, B. (2008). *De l'incomplétude lexicale en traduction automatique : vers une approche morphosémantique multilingue*. Université de Genève.

CARTONI, B. (2009). Lexical morphology in machine translation : A feasibility study. *In 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Athens*, pages 130–138.

COTTEZ, H. (1982). *Dictionnaire des structures du vocabulaire savant*. Les usuels du Robert, Paris.

DAILLE, B. (1999). Identification des adjectifs relationnels en corpus. *In Actes de la Conférence de Traitement Automatique du Langage Naturel (TALN '99)*.

DAILLE, B. (2000). Morphological rule induction for terminology acquisition. *In 18th International Conference on Computational Linguistics (COLING)*, pages 215–221.

DUBOIS, J. et DUBOIS-CHARLIER, F. (1999). *La dérivation suffixale en français*. Nathan Université.

FRUNZA, O. et INKPEN, D. (2010). Word variant identification in old french. *International Journal of Linguistics*, 130:481–510.

GOES, J. (1999). *L'adjectif entre nom et verbe*. De Boeck and Larcier Département Duculot.

HAUER, B. et KONDRAK, G. (2011). Clustering semantically equivalent words into cognate sets in multilingual lists. *In The 5th International Joint Conference on Natural Language Processing (IJCNLP),Chiang Mai*, pages 865–873.

KETKAR, N. S. et YOUNGBLOOD, G. M. (2010). A largest common subsequence-based distance measure for classifying player motion traces in virtual worlds. *In FLAIRS Conference*.

MANIEZ, F. (2005). Identification automatique des adjectifs relationnels : une étude sur corpus. *In De la mesure dans les terme : Presses Universitaires de Lyon*.

MORIN, E. et DAILLE, B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1-2):79–95.

NAMER, F. (2003). Automatiser l'analyse morphosémantique non affixale : le système DériF. *Cahiers de Grammaire*, 28:31–48.

NOAILLY, M. (1999). *L'adjectif en français*. Editions Ophrys.

PEDREIRA, N. R. (2002). De la grammaire traditionnelle à la morphologie dérivationnelle : retour sur l'adjectif de relation. *In VERBA*, pages 421–434.

RAPP, R. (1995). Identifying word translations in non-parallel texts. *In Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL '95), Cambridge, Massachusetts*, pages 320–322.

ROCHETEAU, J. et DAILLE, B. (2011). TTC TermSuite : A UIMA Application for Multilingual Terminology extraction from Comparable Corpora. *In the 5th International Joint Conference on Natural Language Processing (IJCNLP),Chiang Mai*, pages 9–12.

ROCHÉ, M. (2006). Comment les adjectifs sont sémantiquement construits. *Cahier de Grammaire 30*.